Mandoline: Model Evaluation under Distribution Shift

Mayee Chen*, Karan Goel*, Nimit Sohoni*, Fait Poms, Kayvon Fatahalian, Christopher Ré

Stanford University

Motivation

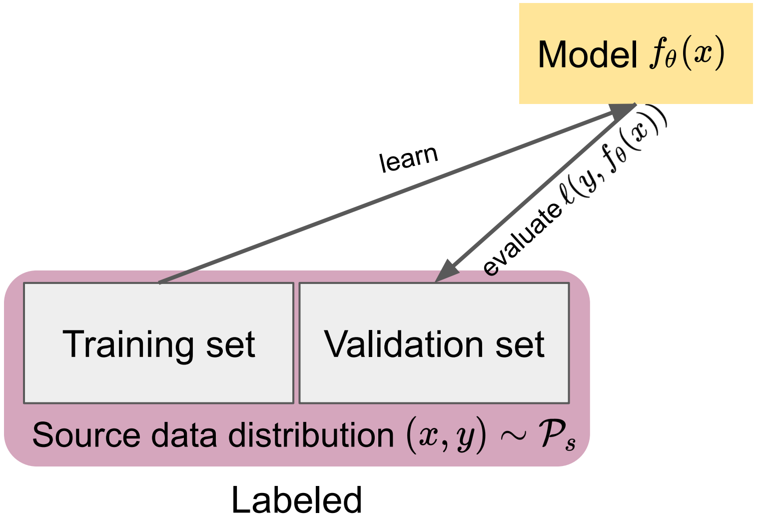
Q: How do we **evaluate** model performance during deployment?

- Model’s deployment setting ≠ training setting

Motivation

Q: How do we **evaluate** model performance during deployment?

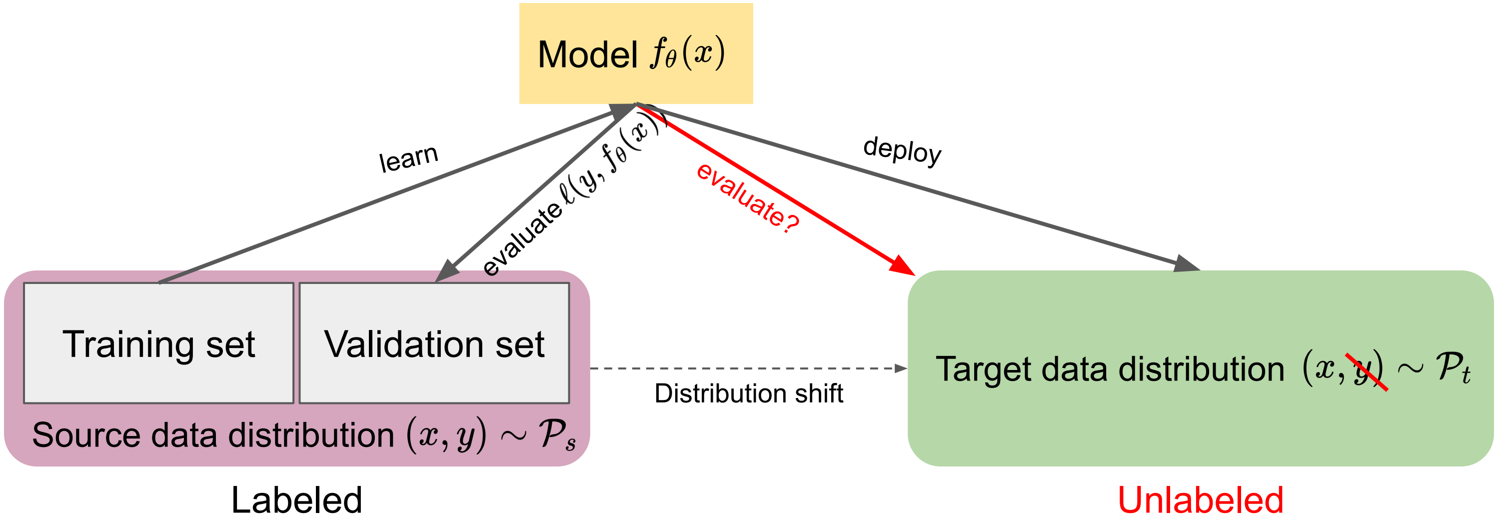
- Model's deployment setting ≠ training setting



Motivation

Q: How do we **evaluate** model performance during deployment?

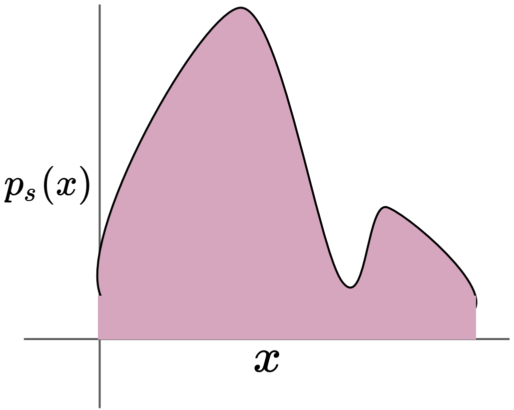
- Model's deployment setting ≠ training setting



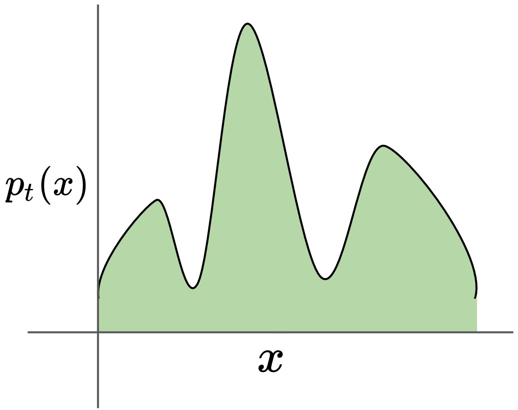
Mandoline: user-guided framework for evaluation under distribution shift

Common approach: importance weighting

Source data distribution $\mathcal{P}_s$



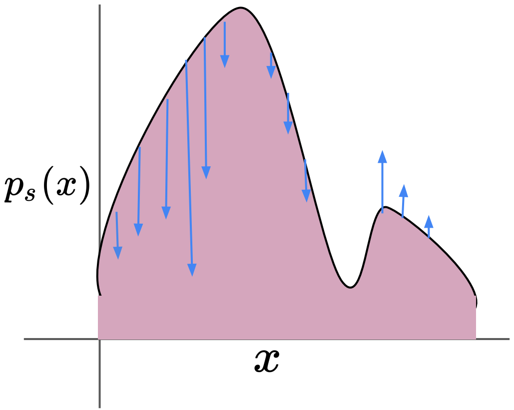
Target data distribution $\mathcal{P}_t$



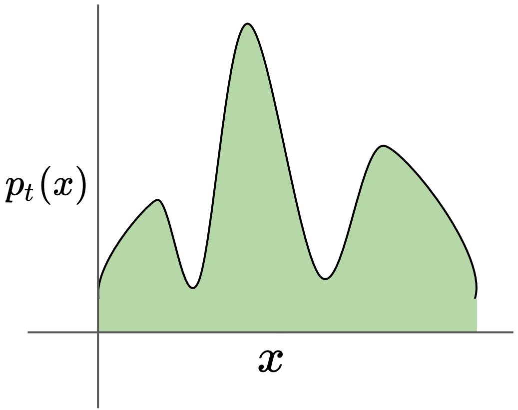
$$\mathbb{E}_t[\ell(y, f_\theta(x))] = \mathbb{E}_s\left[\frac{p_t(x)}{p_s(x)}\ell(y, f_\theta(x))\right] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p_t(x_i)}{p_s(x_i)}\ell(y_i, f_\theta(x_i))$$

Common approach: importance weighting

Source data distribution $\mathcal{P}_s$



Target data distribution $\mathcal{P}_t$

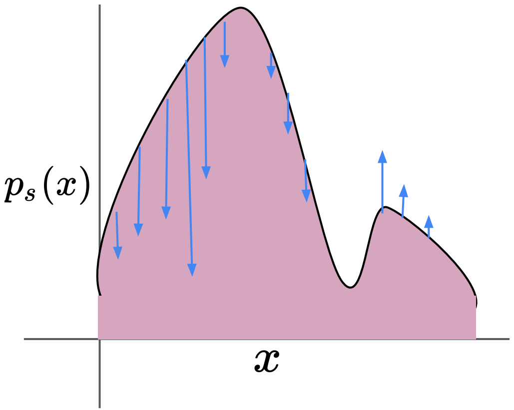


$$\mathbb{E}_t[\ell(y, f_\theta(x))] = \mathbb{E}_s\left[\frac{p_t(x)}{p_s(x)}\ell(y, f_\theta(x))\right] \approx \frac{1}{n}\sum_{i=1}^{n}\boxed{\frac{p_t(x_i)}{p_s(x_i)}}\ell(y_i, f_\theta(x_i))$$

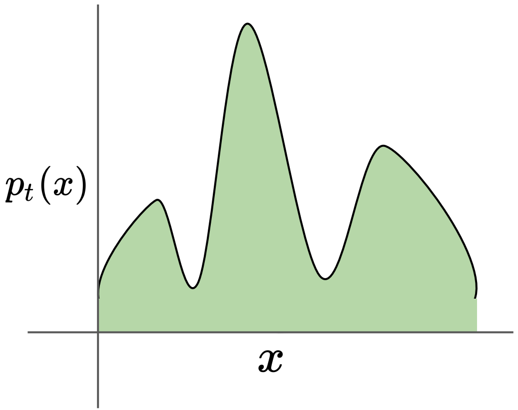
Density ratio to estimate

Common approach: importance weighting

Source data distribution $\mathcal{P}_s$



Target data distribution $\mathcal{P}_t$



$$\mathbb{E}_t[\ell(y, f_\theta(x))] = \mathbb{E}_s\left[\frac{p_t(x)}{p_s(x)}\ell(y, f_\theta(x))\right] \approx \frac{1}{n}\sum_{i=1}^{n} \boxed{\frac{p_t(x_i)}{p_s(x_i)}} \ell(y_i, f_\theta(x_i))$$

Density ratio to estimate

Problems:

- Support shift - what if $p_s(x) = 0, p_t(x) \neq 0$?
- High dimensional data $x \in \mathbb{R}^d$: harder to compute $\frac{p_t(x)}{p_s(x)}$

Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$

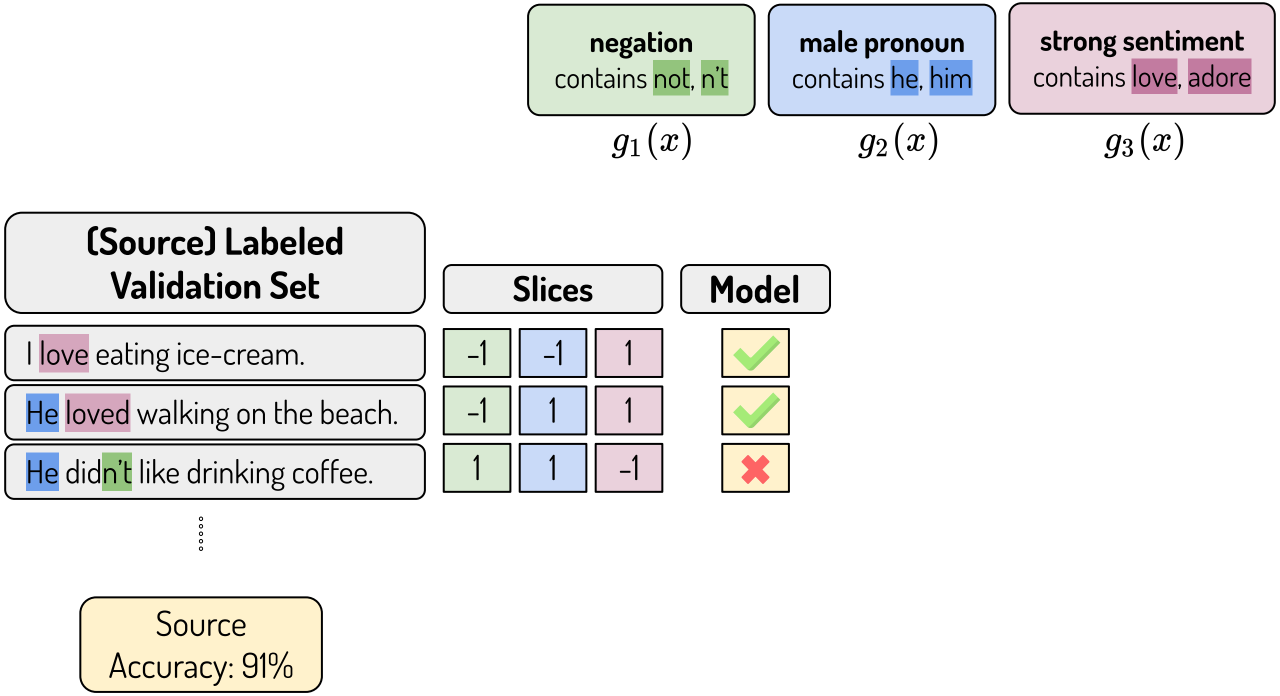
Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$

negation contains not, n't	**male pronoun** contains he, him	**strong sentiment** contains love, adore
$g_1(x)$	$g_2(x)$	$g_3(x)$

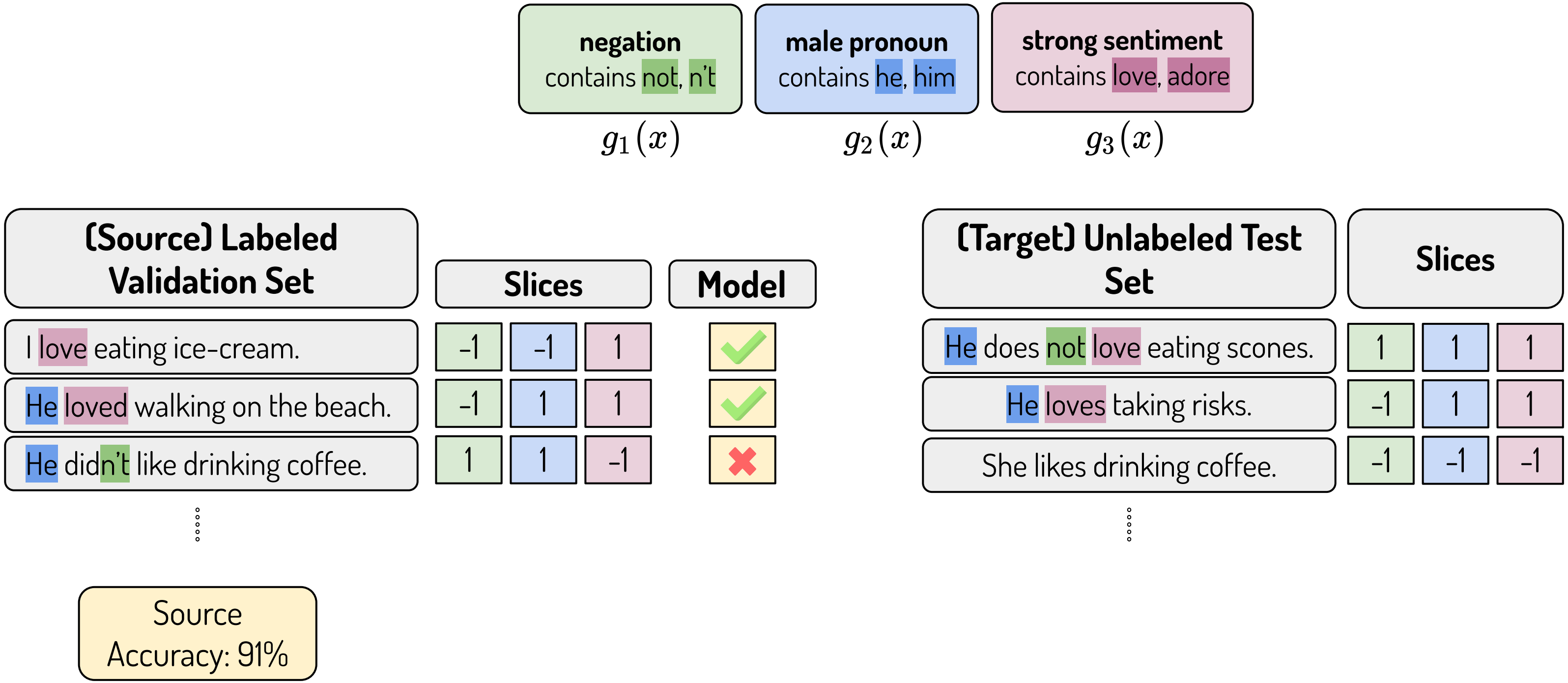
Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$



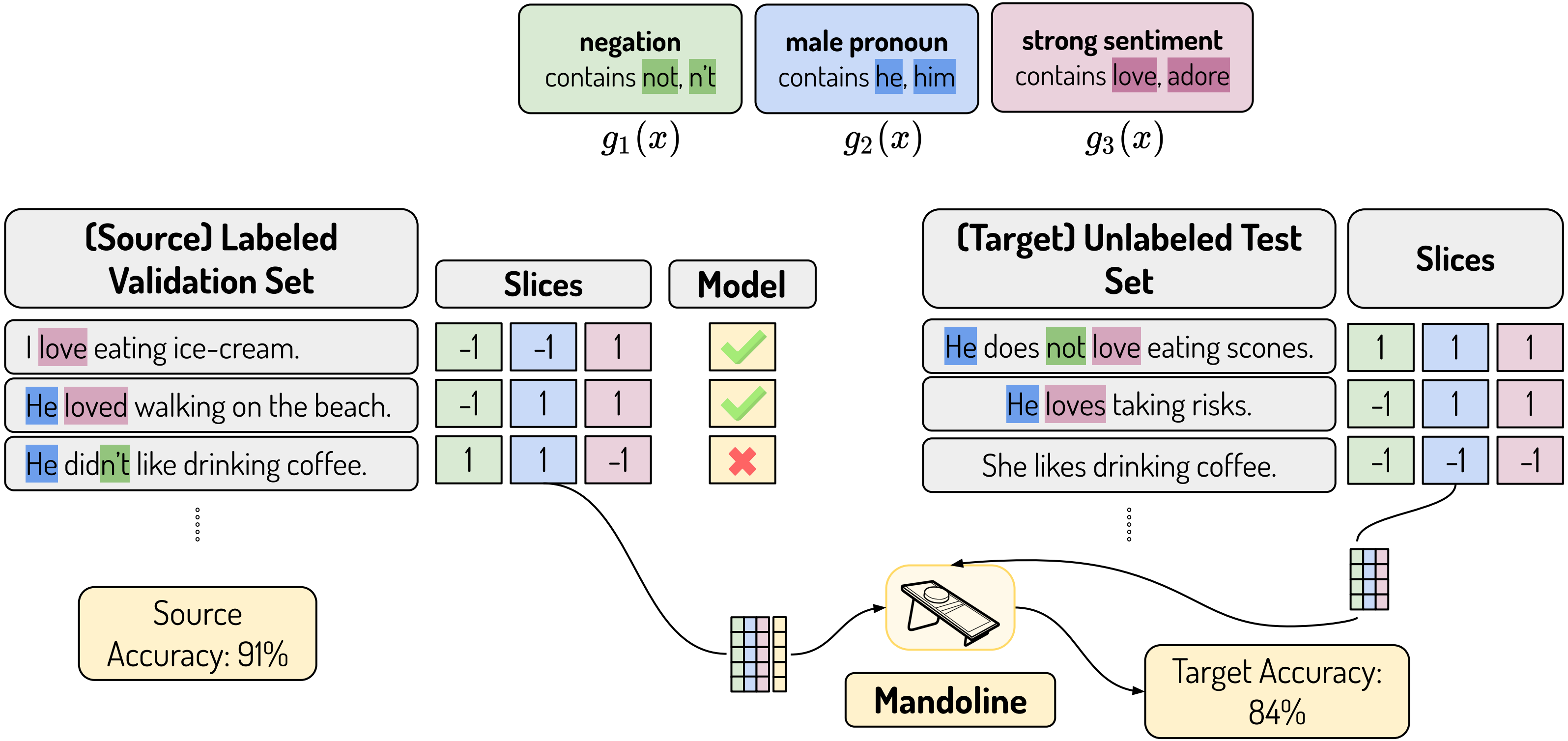
Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$



Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$



Results

Prop 1: if *k* slices $g = \{g_1, \ldots, g_k\}$ *capture all "relevant" distributional shift* between $\mathcal{P}_s$ and $\mathcal{P}_t$, then reweighting with $\frac{p_t(g(x))}{p_s(g(x))}$ recovers $\mathbb{E}_t[\ell(y, f_\theta(x))]$.

- If support shift occurs on irrelevant slices (i.e. slices independent of *Y*), it can be corrected!
- Dimensionality: reduce from $d \rightarrow k$

Results

Prop 1: if *k* slices $g = \{g_1, \ldots, g_k\}$ *capture all "relevant" distributional shift* between $\mathcal{P}_s$ and $\mathcal{P}_t$, then reweighting with $\frac{p_t(g(x))}{p_s(g(x))}$ recovers $\mathbb{E}_t[\ell(y, f_\theta(x))]$.

- If support shift occurs on irrelevant slices (i.e. slices independent of *Y*), it can be corrected!
- Dimensionality: reduce from $d \to k$

- How to compute $\frac{p_t(g(x))}{p_s(g(x))}$? Use any density ratio estimation method on *g(x)*
  - **Kullback-Leibler Importance Estimation Procedure (KLIEP)**
    - Extend to correct for noisily-defined, incomplete slices

Results

Task	Task Labels	Distribution Shift	Slices
CELEBA *image classification*	*male* *vs. female*	$\uparrow$ blurry images	METADATA LABELS *blurry / not blurry*
SNLI→MNLI *natural language inference*	*entailment, neutral or contradiction*	single-genre → multi-genre examples	PROGRAMMATIC *task model predictions, task model entropy*

Results

Task	Task Labels	Distribution Shift	Slices
CelebA *image classification*	*male* *vs. female*	↑ blurry images	METADATA LABELS *blurry / not blurry*
SNLI→MNLI *natural language inference*	*entailment, neutral or contradiction*	single-genre → multi-genre examples	PROGRAMMATIC *task model predictions, task model entropy*

Average Estimation Error (%) Method	CelebA ResNet18	CelebA ResNet50
Source	1.96%	1.74%
CBIW	0.47%	0.53%
KMM	1.97%	1.76%
uLSIF	1.97%	1.76%
Mandoline	**0.16%**	**0.16%**

CelebA

Importance weighting on x

On g(x)

Method	Standard Accuracy Avg. Error	Standard Accuracy Max. Error
Source	$6.2\% \pm 3.8\%$	15.6%
CBIW	$5.5\% \pm 4.5\%$	17.9%
KMM	$5.7\% \pm 3.6\%$	14.6%
uLSIF	$6.4\% \pm 3.9\%$	16.0%
Mandoline	$\mathbf{3.6\% \pm 1.6\%}$	**5.9%**

SNLI → MNLI

Summary

Model evaluation under distribution shift:

- When user-specified slices capture relevant distribution shift, can reweight using them
- Can mitigate 1) support shift and 2) high dimensionality in standard importance reweighting
- Future steps: slice design - frameworks for how to construct good *g*?






Thank you!

Contact: *mfchen@stanford.edu*